24 Minimax Lower Bounds for Stochastic Linear Bandits

Lower bounds for linear bandits turn out to be more nuanced than those for the classical finite-armed bandit. The difference is that for linear bandits the shape of the action set plays a role in the form of the regret, not just the distribution of the noise. This should not come as a big surprise because the stochastic finite-armed bandit problem can be modeled as a linear bandit with actions being the standard basis vectors, $\mathcal{A} = \{e_1, \dots, e_K\}$. In this case the actions are orthogonal, which means that samples from one action do not give information about the rewards for other actions. Other action sets such as the sphere ($\mathcal{A} = S^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$) do not share this property. For example, if $d = 2$ and $\mathcal{A} = S^1$ and an algorithm chooses actions $e_1 = (1, 0)$ and $e_2 = (0, 1)$ many times, then it can deduce the reward it would obtain from choosing any other action.

All results of this chapter have a worst-case flavor showing what is (not) achievable in general, or under a sparsity constraint, or if the realizable assumption is not satisfied. The analysis uses the information-theoretic tools introduced in Part IV combined with careful choices of action sets. The hard part is guessing what is the worst case, which is followed by simply turning the crank on the usual machinery.

In all lower bounds we use a simple model with Gaussian noise. For action $A_t \in \mathcal{A} \subseteq \mathbb{R}^d$ the reward is $X_t = \mu(A_t) + \eta_t$ where $\eta_t \sim \mathcal{N}(0, 1)$ is a sequence of independent standard Gaussian noise and $\mu : \mathcal{A} \to [0, 1]$ is the mean reward. We will usually assume there exists a $\theta \in \mathbb{R}^d$ such that $\mu(a) = \langle a, \theta \rangle$. We write $\mathbb{P}_\mu$ to indicate the measure on outcomes induced by the interaction of the fixed policy and the Gaussian bandit paramterised by μ. Because we are now proving lower bounds it becomes necessary to be explicit about the dependence of the regret on $\mathcal{A}$ and μ or θ. The regret of a policy is:

$$R_n(\mathcal{A}, \mu) = n \max_{a \in \mathcal{A}} \mu(a) - \mathbb{E}_\mu \left[\sum_{t=1}^{n} X_t \right] ,$$

where the expectation is taken with respect to $\mathbb{P}_\mu$. Except in Section 24.4 we assume the reward function is linear, which means there exists a $\theta \in \mathbb{R}^d$ such that $\mu(a) = \langle a, \theta \rangle$. In these cases we write $R_n(\mathcal{A}, \theta)$ and $\mathbb{E}_\theta$ and $\mathbb{P}_\theta$. Recall the notation used for finite-armed bandits by defining $T_x(t) = \sum_{s=1}^{t} \mathbb{I}\{A_s = x\}$.

24.1 Hypercube

The first lower bound is for the hypercube action-set and shows that the upper bounds in Chapter 19 cannot be improved in general.

THEOREM 24.1 *Let $\mathcal{A} = [-1,1]^d$ and $\Theta = \{-n^{-1/2}, n^{-1/2}\}^d$. Then for any policy there exists a $\theta \in \Theta$ such that:*

$$R_n(\mathcal{A},\theta) \geq \frac{\exp(-2)}{8} d\sqrt{n}\,.$$

Proof By the relative entropy identity (Lemma 15.1) we have for $\theta, \theta' \in \Theta$ that

$$\mathrm{D}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{1}{2}\sum_{t=1}^n \mathbb{E}_\theta\left[\langle A_t, \theta - \theta'\rangle^2\right]\,. \tag{24.1}$$

For $i \in [d]$ and $\theta \in \Theta$ define

$$p_{\theta i} = \mathbb{P}_\theta\left(\sum_{t=1}^n \mathbb{I}\left\{\mathrm{sign}(A_{ti}) \neq \mathrm{sign}(\theta_i)\right\} \geq n/2\right)\,.$$

Now let $i \in [d]$ and $\theta \in \Theta$ be fixed and let $\theta' = \theta$ except for $\theta'_i = -\theta_i$. Then by the high probability version of Pinsker's inequality (Theorem 14.2) and Eq. (24.1),

$$p_{\theta i} + p_{\theta' i} \geq \frac{1}{2}\exp\left(-\frac{1}{2}\sum_{t=1}^n \mathbb{E}_\theta[\langle A_t, \theta - \theta'\rangle^2]\right) \geq \frac{1}{2}\exp\left(-2\right)\,. \tag{24.2}$$

Applying an 'averaging hammer' over all $\theta \in \Theta$, which satisfies $|\Theta| = 2^d$:

$$\sum_{\theta\in\Theta}\frac{1}{|\Theta|}\sum_{i=1}^d p_{\theta i} = \frac{1}{|\Theta|}\sum_{i=1}^d\sum_{\theta\in\Theta} p_{\theta i} \geq \frac{d}{4}\exp\left(-2\right)\,.$$

Since $p_{\theta i}$ is nonnegative this implies there exists a $\theta \in \Theta$ such that $\sum_{i=1}^d p_{\theta i} \geq d\exp\left(-2\right)/4$. By the definition of $p_{\theta i}$ the regret for this choice of θ is at least

$$\begin{aligned}
R_n(\mathcal{A},\theta) &\geq \sqrt{\frac{1}{n}}\sum_{i=1}^d \mathbb{E}\left[\sum_{t=1}^n \mathbb{I}\left\{\mathrm{sign}(A_{ti}) \neq \mathrm{sign}(\theta_i)\right\}\right] \\
&\geq \frac{\sqrt{n}}{2}\sum_{i=1}^d \mathbb{P}_\theta\left(\sum_{t=1}^n \mathbb{I}\left\{\mathrm{sign}(A_{ti}) \neq \mathrm{sign}(\theta_i)\right\} \geq n/2\right) \\
&= \frac{\sqrt{n}}{2}\sum_{i=1}^d p_{\theta i} \geq \frac{\exp(-2)}{8} d\sqrt{n}\,.
\end{aligned}$$

□

Except for logarithmic factors this shows the algorithm of Chapter 19 is near-optimal for this action set. The same proof works when $\mathcal{A} = \{-1,1\}^d$ is restricted to the corners of the hypercube, which is a finite-armed linear bandit. In Chapter 22 we gave a policy with regret $R_n = O(\sqrt{nd\log(nK)})$ where

$K = |\mathcal{A}|$. There is no contradiction because the action set in the above proof has $K = |\mathcal{A}| = 2^d$.

24.2 Sphere

Lower bounding the minimax regret when the action-set is the sphere presents an additional challenge relative to the hypercube. The product structure of the hypercube means that the learner can treat the dimensions independently, which is reflected in the lower bound. For the sphere this is not true because the magnitude of the action in one dimension constrains the learner in other dimensions. Nevertheless, almost the same technique with one modification allows us to prove a similar bound.

THEOREM 24.2 *Assume $d \le 2n$ and let $\mathcal{A} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$. Then there exists a $\theta \in \mathbb{R}^d$ with $\|\theta\|_2^2 = d^2/(4n)$ such that $R_n(\mathcal{A}, \theta) \ge d\sqrt{n}/16$.*

Proof Let $\Delta = \frac{1}{4}\sqrt{d/n}$ and $\theta \in \{\pm\Delta\}^d$ and $\tau_i = n \wedge \min\{t : \sum_{s=1}^t A_{si}^2 \ge n/d\}$. Then

$$\begin{aligned} R_n(\mathcal{A}, \theta) &= \Delta \mathbb{E}_\theta \left[\sum_{t=1}^n \sum_{i=1}^d \left(\frac{1}{\sqrt{d}} - A_{ti} \operatorname{sign}(\theta_i) \right) \right] \\ &= \frac{\Delta\sqrt{d}}{2} \mathbb{E}_\theta \left[\sum_{t=1}^n \sum_{i=1}^d \left(\frac{1}{\sqrt{d}} - A_{ti} \operatorname{sign}(\theta_i) \right)^2 \right] \\ &\ge \frac{\Delta\sqrt{d}}{2} \sum_{i=1}^d \mathbb{E}_\theta \left[\sum_{t=1}^{\tau_i} \left(\frac{1}{\sqrt{d}} - A_{ti} \operatorname{sign}(\theta_i) \right)^2 \right] . \end{aligned}$$

Let $U_i(x) = \sum_{t=1}^{\tau_i} (1/\sqrt{d} - A_{ti}x)^2$ and $\theta' \in \{\pm\Delta\}^d$ be another parameter vector such that $\theta_j = \theta'_j$ for $j \ne i$ and $\theta'_i = -\theta_i$ and assume without loss of generality that $\theta_i > 0$. Let $\mathbb{P}$ and $\mathbb{P}'$ be the laws of $U_i(1)$ with respect to the bandit/learner interaction measure induced by θ and θ' respectively, then

$$\begin{aligned} \mathbb{E}_\theta[U_i(1)] &\ge \mathbb{E}_{\theta'}[U_i(1)] - \left(\frac{4n}{d} + 2 \right) \sqrt{\frac{1}{2} \mathrm{D}(\mathbb{P}, \mathbb{P}')} \\ &\ge \mathbb{E}_{\theta'}[U_i(1)] - \frac{\Delta}{2} \left(\frac{4n}{d} + 2 \right) \sqrt{\mathbb{E}\left[\sum_{t=1}^{\tau_i} A_{ti}^2 \right]} \\ &\ge \mathbb{E}_{\theta'}[U_i(1)] - \frac{\Delta}{2} \left(\frac{4n}{d} + 2 \right) \sqrt{\frac{n}{d}} \\ &\ge \mathbb{E}_{\theta'}[U_i(1)] - \frac{4\Delta n}{d} \sqrt{\frac{n}{d}} , \end{aligned}$$

where in the first inequality we used Pinsker's inequality (Eq. (14.9)) and the

bound $U_i(1) \le 4n/d + 2$, which follows from the definition of τ_i and the fact that $|A_{\tau_i i}| \le 1$. In the second line we used the chain rule for the relative entropy up to a stopping time (Exercise 15.6). The second last inequality is true by the definition of τ_i and the last by the assumption that $d \le 2n$.

$$\begin{aligned}\mathbb{E}_\theta[U_i(1)] + \mathbb{E}_{\theta'}[U_i(-1)] &\ge \mathbb{E}_{\theta'}[U_i(1) + U_i(-1)] - \frac{4n\Delta}{d}\sqrt{\frac{n}{d}} \\ &= 2\mathbb{E}_{\theta'}\left[\frac{\tau_i}{d} + \sum_{t=1}^{\tau_i} A_{ti}^2\right] - \frac{4n\Delta}{d}\sqrt{\frac{n}{d}} \ge \frac{2n}{d} - \frac{4n\Delta}{d}\sqrt{\frac{n}{d}} = \frac{n}{d}\,.\end{aligned}$$

The proof is completed using the randomization hammer:

$$\begin{aligned}\sum_{\theta \in \{\pm\Delta\}^d} R_n(\mathcal{A}, \theta) &\ge \frac{\Delta\sqrt{d}}{2} \sum_{i=1}^{d} \sum_{\theta \in \{\pm\Delta\}^d} \mathbb{E}_\theta[U_i(\operatorname{sign}(\theta_i))] \\ &= \frac{\Delta\sqrt{d}}{2} \sum_{i=1}^{d} \sum_{\theta_{-i} \in \{\pm\Delta\}^{d-1}} \sum_{\theta_i \in \{\pm\Delta\}} \mathbb{E}_\theta[U_i(\operatorname{sign}(\theta_i))] \\ &\ge \frac{\Delta\sqrt{d}}{2} \sum_{i=1}^{d} \sum_{\theta_{-i} \in \{\pm\Delta\}^{d-1}} \frac{n}{d} = 2^{d-2} n\Delta\sqrt{d}\,.\end{aligned}$$

Hence there exists a $\theta \in \{\pm\Delta\}^d$ such that $R_n(\mathcal{A}, \theta) \ge \dfrac{n\Delta\sqrt{d}}{4} = \dfrac{d\sqrt{n}}{16}$. □

24.3 Sparse parameter vectors

In Chapter 23 we gave an algorithm with $R_n = \tilde{O}(\sqrt{dpn})$ where $p \ge \|\theta\|_0$ is a known bound on the sparsity of the unknown parameter. Except for logarithmic terms this bound cannot be improved. An extreme case is when $p = 1$, which essentially reduces to the finite-armed bandit problem where the minimax regret has order $\sqrt{dn}$ (see Chapter 15). For this reason we cannot expect too much from sparsity and in particular the worst case bound will depend on polynomially on the ambient dimension d.

Constructing a lower bound for $p > 1$ is relatively straightforward. For simplicity we assume that $d = pk$ for some integer $k > 1$. A sparse linear bandit can mimic the learner playing p finite-armed bandits simultaneously, each with k arms. Rather than observing the reward for each bandit, however, the learner only observes the sum of the rewards and the noise is added at the end. This is sometimes called the **multitask bandit** problem.

THEOREM 24.3 *Assume $pd \le n$ and there exists a natural number $k > 1$ such that $d = pk$. Let $\mathcal{A} = \{e_i : i \in [k]\}^p \subset \mathbb{R}^d$. Then for any policy there exists a $\theta \in \mathbb{R}^d$ with $\|\theta\|_0 = p$ and $\|\theta\|_\infty \le \sqrt{d/(pn)}$ such that $R_n(\mathcal{A}, \theta) \ge \frac{1}{8}\sqrt{pdn}$.*

Proof Let $\Delta > 0$ and $\Theta = \{\Delta e_i : i \in [k]\} \subset \mathbb{R}^k$. Given $\theta \in \Theta^p$ and $i \in [p]$ let $\theta^{(i)} \in \mathbb{R}^k$ be defined by $\theta^{(i)}_k = \theta_{(i-1)p+k}$, which means that

$$\theta^\top = [\theta^{(1)\top}, \theta^{(2)\top}, \dots, \theta^{(p)\top}] .$$

Next define matrix $V \in \mathbb{R}^{p \times d}$ be the matrix with $V_{ij} = 1 + (j-1) \bmod k$. For example, when $p = 2$:

$$V = \begin{bmatrix} 1 & \cdots & k & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & k & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & \cdots & k \end{bmatrix} .$$

Let $B_t = VA_t \in [k]^p$ represent the vector of 'base' actions chosen by the learner in each of the p bandits in round t. The optimal action in the ith bandit is

$$b^*_i(\theta) = \operatorname{argmax}_{b \in [k]} \theta^{(i)}_b .$$

The regret can be decomposed into

$$R_n(\theta) = \sum_{i=1}^{p} \underbrace{\Delta \mathbb{E}_\theta \left[\sum_{t=1}^{n} \mathbb{I}\{B_{ti} \neq b^*_i\} \right]}_{R_{ni}(\theta)} .$$

For $i \in [p]$ we abbreviate $\theta^{(-i)} = \theta^{(1)}, \dots, \theta^{(i-1)}, \theta^{(i+1)}, \cdots, \theta^{(p)}$. Then

$$\begin{aligned} \frac{1}{|\Theta|^p} \sum_{\theta \in \Theta^p} R_n(\theta) &= \frac{1}{|\Theta|^p} \sum_{i=1}^{p} R_{ni}(\theta) \\ &= \sum_{i=1}^{p} \frac{1}{|\Theta|^{p-1}} \sum_{\theta^{(-i)} \in \Theta^{p-1}} \frac{1}{|\Theta|} \sum_{\theta^{(i)} \in \Theta} R_{ni}(\theta) \\ &\geq \frac{1}{8} \sum_{i=1}^{p} \frac{1}{|\Theta|^{p-1}} \sum_{\theta^{(-i)} \in \Theta^{p-1}} \sqrt{kn} \qquad (24.3) \\ &= \frac{1}{8} p \sqrt{kn} = \frac{1}{8} \sqrt{dpn} . \end{aligned}$$

The only tricky step is the inequality, which follows by choosing $\Delta \approx \sqrt{k/n}$ and repeating the argument outlined in Exercise 15.1. We leave it to the reader to check the details (Exercise 24.1). □

24.4 Unrealizable case

An important generalization of the linear model is the **unrealizable** case where the mean rewards are not assumed to follow a linear model exactly. Suppose that $\mathcal{A} \subset \mathbb{R}^d$ is a finite set with $|\mathcal{A}| = K$ and that $X_t = \eta_t + \mu(A_t)$ where $\mu : \mathcal{A} \to \mathbb{R}$ is an unknown function. Let $\theta \in \mathbb{R}^d$ be the parameter vector for which

$\sup_{a\in\mathcal{A}} |\langle\theta, a\rangle - \mu(a)|$ is as small as possible:

$$\theta = \operatorname{argmin}_{\theta\in\mathbb{R}^d} \sup_{a\in\mathcal{A}} |\langle\theta, a\rangle - \mu(a)| \,.$$

Then let $\varepsilon = \sup_{a\in\mathcal{A}} |\langle\theta, a\rangle - \mu(a)|$ be the maximum error. It would be very pleasant to have an algorithm such that

$$R_n(\mathcal{A}, \mu) = n \max_{a\in\mathcal{A}} \mu(a) - \mathbb{E}\left[\sum_{t=1}^n \mu(A_t)\right] = \tilde{O}(\min\{d\sqrt{n} + \varepsilon n,\, \sqrt{Kn}\}) \,. \quad (24.4)$$

Unfortunately it turns out that results of this kind are not achievable. To show this we will prove a generic bound for the classical finite-armed bandit problem and afterwards show how this implies the impossibility of an adaptive bound like the above.

THEOREM 24.4 *Let $\mathcal{A} = [K]$ and for $\mu \in [0,1]^K$ the reward is $X_t = \mu_{A_t} + \eta_t$ and the regret is*

$$R_n(\mu) = n \max_{i\in\mathcal{A}} \mu_i - \mathbb{E}_\mu\left[\sum_{t=1}^n \mu_{A_t}\right] \,.$$

Define $\Theta, \Theta' \subset \mathbb{R}^K$ by

$$\Theta = \left\{\mu \in [0,1]^K : \mu_i = 0 \text{ for } i > 1\right\} \qquad \Theta' = \left\{\mu \in [0,1]^K\right\} \,.$$

If $V \in \mathbb{R}$ is such that $2(K-1) \le V \le \sqrt{n(K-1)\exp(-2)/8}$ and $\sup_{\mu\in\Theta} R_n(\mu) \le V$, then

$$\sup_{\mu'\in\Theta'} R_n(\mu') \ge \frac{n(K-1)}{8V} \exp(-2) \,.$$

Proof Recall that $T_i(n) = \sum_{t=1}^n \mathbb{I}\{A_t = i\}$ is the number of times arm i is played after all n rounds. Let $\mu \in \Theta$ be given by $\mu_1 = \Delta = (K-1)/V \le 1/2$. The regret is then decomposed as:

$$R_n(\mu) = \Delta \sum_{i=2}^K \mathbb{E}_\mu[T_i(n)] \le V \,.$$

Rearranging shows that $\sum_{i=2}^K \mathbb{E}_\mu[T_i(n)] \le \frac{V}{\Delta}$ and so by the pigeonhole principle there exists an $i > 1$ such that

$$\mathbb{E}_\mu[T_i(n)] \le \frac{V}{(K-1)\Delta} = \frac{1}{\Delta^2} \,.$$

Then define $\mu' \in \Theta'$ by

$$\mu'_j = \begin{cases} \Delta & \text{if } j = 1 \\ 2\Delta & \text{if } j = i \\ 0 & \text{otherwise} \,. \end{cases}$$

Then by Theorem 14.2 and Lemma 15.1, for any event A we have

$$\mathbb{P}_\mu(A) + \mathbb{P}_{\mu'}(A^c) \geq \frac{1}{2} \exp\left(\mathrm{D}(\mathbb{P}_\mu, \mathbb{P}_{\mu'})\right) = \frac{1}{2} \exp\left(-2\Delta^2 \mathbb{E}[T_i(n)]\right) \geq \frac{1}{2} \exp\left(-2\right) .$$

By choosing $A = \{T_1(n) \leq n/2\}$ we have

$$R_n(\mu) + R_n(\mu') \geq \frac{n\Delta}{4} \exp(-2) = \frac{n(K-1)}{4V} \exp(-2) .$$

Therefore by the assumption that $R_n(\mu) \leq V \leq \sqrt{n(K-1)\exp(-2)/8}$ we have

$$R_n(\mu') \geq \frac{n(K-1)}{8V} \exp(-2) . \qquad \square$$

As promised we now relate this to the unrealizable linear bandits. Suppose that $d = 1$ (an absurd case) and that there are K arms $\mathcal{A} = \{a_1, a_2, \ldots, a_K\} \subset \mathbb{R}^1$ where $a_1 = (1)$ and $a_i = (0)$ for $i > 1$. Clearly if $\theta > 0$ and $\mu(a_i) = \langle a_i, \theta \rangle$, then the problem can be modelled as a finite-armed bandit with means $\mu \in \Theta \subset [0,1]^K$. In the general case we just have a finite-armed bandit with $\mu \in \Theta'$. If in the first case we have $R_n(\mathcal{A}, \mu) = O(\sqrt{n})$, then the theorem shows for large enough n that

$$\sup_{\mu \in \Theta'} R_n(\mathcal{A}, \mu) = O(K\sqrt{n}) .$$

It follows that Eq. (24.4) is a pipe dream. To our knowledge it is still an open question of what is possible on this front. Our conjecture is that there is a policy for which

$$R_n(\mathcal{A}, \theta) = \tilde{O}\left(\min\left\{d\sqrt{n} + \varepsilon n, \frac{K}{d}\sqrt{n}\right\}\right) .$$

In fact, it is not hard to design an algorithm that tries to achieve this bound by assuming the problem is realizable, but using some additional time to explore the remaining arms up to some accuracy to confirm the hypothesis.

24.5 Notes

1 The worst-case bound demonstrates the near-optimality of the OFUL algorithm for a specific action set. It is an open question to characterize the optimal regret for a wide range of action sets. We will return to these issues soon when we discuss adversarial linear bandits.

24.6 Bibliographic remarks

Worst-case lower bounds for stochastic bandits have appeared in a variety of places, all with roughly the same bound, but for different action sets. Our very simple proof for the hypercube is new, but takes inspiration from the paper by Shamir [2015]. The first lower bound for the sphere was given by Rusmevichientong and

Tsitsiklis [2010] with smaller constants and a complicated proof. As far as we know the first lower bound of $\Omega(d\sqrt{n})$ was given by Dani et al. [2008] for an action-set equal to the product of 2-dimensional disks. The results for the unrealizable case are inspired by the work of one of the authors on the Pareto-regret frontier for bandits, which characterizes what trade-offs are available when it is desirable to have a regret that is unusually small relative to some specific arms [Lattimore, 2015a].

24.7 Exercises

24.1 Completing the missing steps to prove the inequality in Eq. (24.3).